NVM Express Technical Errata

Errata ID	**006**
Affected Spec Ver.	**NVM Express 1.0**
Corrected Spec Ver.	

Submission info

Name	Company	Date
Santosh	Samsung	3/22/2011
Kevin Marks	Dell	3/22/2011
Amber Huffman	Intel	3/22/2011
Mark Schmisseur	Intel	3/22/2011

This erratum corrects several items in the specification; most are editorial.

The erratum adds clarifications to the command set sections describing where the Submission Queue and Completion Queue entries are defined.

The erratum corrects that the Get Log Page command is Namespace specific.

The Interrupt Mask Set and Interrupt Mask Clear register language was clarified on reads.

It is clarified that metadata shall be written atomically with its associated LBA.

Several editorial changes are made to section 2.

Description of the specification technical flaw

Modify the first three paragraphs of section 5 as shown below:

5 Admin Command Set

The Admin Command Set defines the commands that may be issued to the Admin Submission Queue.

The Submission Queue Entry (SQE) structure and the fields that are common to all Admin commands are defined in section 4.2. The Completion Queue Entry (CQE) structure and the fields that are common to all Admin commands are defined in section 4.5. The command specific fields in the SQE and CQE structures for the Admin Command Set are defined in this section.

For all Admin commands, Dword 14 and 15 are I/O Command Set specific.

Modify the first three paragraphs of section 6 as shown below:

6 NVM Command Set

The device is comprised of some number of namespaces, where each namespace is comprised of some number of logical blocks. A logical block is the smallest unit of data that may be read or written from the controller. The logical block data size, reported in bytes, is always a power of two. LBA sizes may be 512 bytes, 1KB, 2KB, 4KB, 8KB, etc. Supported LBA sizes are reported in the Identify Namespace data structure.

The NVM command set includes the commands listed in Figure 98. The following subsections describe the definition for each of these commands. Commands shall only be issued by the host when the controller is ready as indicated in the Controller Status register (CSTS.RDY) and after appropriate I/O Submission Queue(s) and I/O Completion Queue(s) have been created.

The Submission Queue Entry (SQE) structure and the fields that are common to all NVM commands are defined in section 4.2. The Completion Queue Entry (CQE) structure and the fields that are common to all NVM commands are defined in section 4.5. The command specific fields in the SQE and CQE structures for the NVM Command Set are defined in this section.

In the case of Compare, Read, and Write commands, the host may indicate whether a time limit should be applied to the operation by setting the Limited Retry (LR) field in the command. The time limit is indicated in the Error Recovery feature, specified in section 5.12.1.5. If the host does not indicate a time limit should be applied, then the controller should apply all error recovery means to complete the operation.

Modify Figure 24 as shown below:

Figure 24: Opcodes for Admin Commands

Opcode (07) Generic Command	Opcode (06:02) Function	Opcode (01:00) Data Transfer	Opcode	O/M	Namespace Identifier Used[3]	Command
0b	000 00b	00b	00h	M	No	Delete I/O Submission Queue
0b	000 00b	01b	01h	M	No	Create I/O Submission Queue
0b	000 00b	10b	02h	M	~~No~~ Yes	Get Log Page
0b	000 01b	00b	04h	M	No	Delete I/O Completion Queue
0b	000 01b	01b	05h	M	No	Create I/O Completion Queue
0b	000 01b	10b	06h	M	Yes	Identify
0b	000 10b	00b	08h	M	No	Abort
0b	000 10b	01b	09h	M	Yes	Set Features
0b	000 00b	10b	0Ah	M	Yes	Get Features
0b	000 11b	00b	0Ch	M	No	Asynchronous Event Request
0b	001 00b	00b	10h	O	No	Firmware Activate
0b	001 00b	01b	11h	O	No	Firmware Image Download
	I/O Command Set Specific					
1b	na	na	80h – BFh	O		I/O Command Set specific
	Vendor Specific					
1b	na	na	C0h – FFh	O		Vendor specific

NOTES:
1. O/M definition: O = Optional, M = Mandatory.
2. Opcodes not listed are reserved.
3. A subset of commands uses the Namespace Identifier field (CDW1.NSID). When not used, the field shall be cleared to 0h. For the Get Features and Set Features command, the Namespace Identifier is only used for the LBA Range Type feature. For the Identify command, the Namespace Identifier is only used for the Namespace data structure. For the Get Log Page command, a value of FFFFFFFFh is used to specify that the global values should be returned.

Modify the register definition in section 3.1.3 as shown below:

Bit	Type	Reset	Description
31:00	RW1S	0h	**Interrupt Vector Mask Set (IVMS):** This field is bit significant. If a '1' is written to a bit, then the corresponding interrupt vector is masked. Writing a '0' to a bit has no effect. When read, this field returns the current interrupt mask value within the controller (not the value of this register). If a bit has a value of a '1', then the corresponding interrupt vector is masked. If a bit has a value of '0', then the corresponding interrupt vector is not masked.

Modify the register definition in section 3.1.4 as shown below:

Bit	Type	Reset	Description
31:00	RW1C	0h	**Interrupt Vector Mask Clear (IVMC):** This field is bit significant. If a '1' is written to a bit, then the corresponding interrupt vector is unmasked. Writing a '0' to a bit has no effect. When read, this field returns the current interrupt mask value within the controller (not the value of this register). If a bit has a value of a '1', then the corresponding interrupt vector is masked, If a bit has a value of '0', then the corresponding interrupt vector is not masked.

Modify section 4.4 as shown below:

4.4 Metadata Region (MR)

Metadata may be supported for a namespace as either part of the LBA (creating an extended LBA which is a larger LBA that is exposed to the application) or it may be transferred as a separate contiguous buffer of data. The metadata shall not be split between the LBA and a separate metadata buffer. For writes, the metadata shall be written atomically with its associated LBA. Refer to section 8.2.

In the case where the namespace is formatted to transfer the metadata as a separate contiguous buffer of data, then the Metadata Region is used. In this case, the location of the Metadata Region is indicated by the Metadata Pointer within the command. The Metadata Pointer within the command shall be Dword aligned.

The controller may support several physical formats of LBA size and associated metadata size. There may be performance differences between different physical formats. This is indicated as part of the Identify Namespace data structure.

If the namespace is formatted to use end-to-end data protection, then the first eight bytes or last eight bytes of the metadata is used for protection information (specified as part of the NVM Format operation).

Modify section 2.1.9 as shown below:

2.1.9 Offset 0Fh: BIST – Built In Self Test (Optional)

The following register is optional, but if implemented, shall look as follows. When not implemented, it shall be read-only 00h.

Bits	Type	Reset	Description
07	RO	Impl Spec	**BIST Capable (BC):** Indicates whether the controller has a BIST function.
06	RW	0	**Start BIST (SB):** Host software ~~Software~~ sets this bit to '1' to invoke BIST. The controller clears this bit to '0' when BIST is complete.
05:04	RO	00	Reserved
03:00	RO	0h	**Completion Code (CC):** Indicates the completion code status of BIST. A non-zero value indicates a failure.

Modify section 2.1.12 as shown below:

2.1.12 Offset 18h: IDBAR (BAR2) – Index/Data Pair Register Base Address (Optional)

This register specifies the Index/Data Pair base address. These registers are used to access the memory registers defined in section **Error! Reference source not found.** using I/O based accesses. ~~**Note:** This functionality is optional.~~ If Index/Data Pair is not supported, then the IDBAR shall be read only 0h.

Bit	Type	Reset	Description
31:03	RW	0	**Base Address (BA):** Base address of Index/Data Pair registers that is 8 bytes in size.
02:01	RO	0	Reserved
00	RO	1	**Resource Type Indicator (RTE):** Indicates a request for register I/O space.

Modify section 2.1.18 as shown below:

2.1.18 Offset 30h: EROM – Expansion ROM (Optional)

~~The following register is optional.~~ If the register is not implemented, it shall be read-only 00h.

Bit	Type	Reset	Description
31:00	RW	Impl Spec	**ROM Base Address (RBA):** Indicates the base address of the controller's expansion ROM. Not supported for integrated implementations.

Modify section 2.1.20 as shown below:

2.1.20 Offset 3Ch: INTR - Interrupt Information

Bits	Type	Reset	Description
15:08	RO	Impl Spec	**Interrupt Pin (IPIN):** This indicates the interrupt pin the controller uses.
07:00	RW	00h	**Interrupt Line (ILINE):** Host software ~~Software~~ written value to indicate which interrupt line (vector) the interrupt is connected to. No hardware action is taken on this register.

Modify section 2.4.3 as shown below:

2.4.3 Offset MSIXCAP + 4h: MTAB – MSI-X Table Offset / Table BIR

<table>
<tr><th>Bits</th><th>Type</th><th>Reset</th><th>Description</th></tr>
<tr><td>31:03</td><td>RO</td><td>Impl Spec</td><td>Table Offset (TO): Used as an offset from the address contained by one of the function's Base Address registers to point to the base of the MSI-X Table. The lower three Table BIR bits are masked off (cleared to 000b) by system software to form a 32-bit Qword-aligned offset.</td></tr>
<tr><td>02:00</td><td>RO</td><td>Impl Spec</td><td>Table BIR (TBIR): This field indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the function's MSI-X Table into system memory.

<table>
<tr><th>BIR Value</th><th>BAR Offset</th></tr>
<tr><td>0</td><td>10h</td></tr>
<tr><td>1</td><td>na</td></tr>
<tr><td>2</td><td>na</td></tr>
<tr><td>3</td><td>Reserved</td></tr>
<tr><td>4</td><td>20h</td></tr>
<tr><td>5</td><td>24h</td></tr>
<tr><td>6</td><td>Reserved</td></tr>
<tr><td>7</td><td>Reserved</td></tr>
</table>

For a 64-bit Base Address register, the Table BIR indicates the lower Dword. With PCI-to-PCI bridges, BIR values 2 through 5 are also reserved.</td></tr>
</table>

Modify bits 02:00 in the PXDCAP register in section 2.5.3 as shown below:

02:00	RO	Impl Spec	**Max_Payload_Size Supported (MPS):** This field indicates the maximum payload size that the function ~~can~~ may support for TLPs.

Modify bits 14:12 in the PXDC register in section 2.5.4 as shown below:

14:12	RW/ RO	Impl Spec	**Max_Read_Request_Size (MRRS):** This field sets the maximum Read Request size for the Function as a Requester. The Function ~~must~~ shall not generate Read Requests with size exceeding the set value.

Disposition log

3/22/2011	Erratum captured.
3/24/2011	Updates to the INTMS/INTMC definitions.
5/10/2011	Erratum ratified.